



DRAFT
DEPARTMENT OF HEALTH & HUMAN SERVICES

MEMORANDUM

Public Health Service
Food and Drug Administration
1350 Piccard Drive
Rockville, MD 20850

Date: March 3, 2009

From: Ronald Kaczmarek, MD, MPH
Cara J. Krulewitch, CNM, PhD
Nilsa Loyo-Berríos, PhD, MSc
Art Sedrakyan, MD, PhD
Cunlin Wang, MD, PhD
Division of Epidemiology (DEPI), HFZ-541
Office of Surveillance and Biometrics (OSB)

Subject: Epidemiologic Review: CAS Mammography

To: Simon Choi, PhD
Radiologic Products Network, HFZ-4

Through: Danica Marinac-Dabic, MD, PhD, DE/OSB, HFZ-541 _____

Purpose:

The purpose of this memorandum is to present an epidemiologic review of the scientific literature on the use of CADe mammography as part of breast cancer screening programs.

I. Introduction

The detection of breast cancer is a vital public health issue. The National Cancer Institute, employing data from the Surveillance, Epidemiology and End Results Program, has estimated that 182,460 women would be diagnosed with breast cancer in 2008. Fully 40,480 of these women would die from the disease. Mammographic screening has been demonstrated in several clinical trials to decrease the mortality from breast cancer.

Mammography is widely recognized as a particularly challenging examination for radiologists to correctly read. These challenges are amplified by the nature of mammographic screening in actual practice, where less than one percent of all screening mammograms will be true positives. Numerous retrospective studies have confirmed that the review of screening mammograms in women later diagnosed with breast cancer frequently reveals abnormalities that can be detected on earlier examinations. Although the practice of single reader screening mammography is effective in reducing mortality from breast cancer, there can be no doubt that there is considerable room for improvement in breast cancer detection.

Computer-assisted detection (CADe) devices were developed for mammography to assist the

radiologist in the reading of mammographic examinations. CADe devices analyze digitized or digital mammography images using software programs to find features that are associated with breast cancer. A mark is placed at the site of these findings for the radiologist to review. Independent radiologist reading should occur before the provision of CADe input.

II. Methods

A search of the MEDLINE database was performed using the following terms that define CADe mammography and CADe: all MeSH terms for 'mammography', 'CADe mammography', 'Computer assisted detection', 'Computer-aided detection', 'Computer-based detection', 'computer-based diagnosis', 'computer-aided diagnosis', 'computer assisted diagnosis' 'diagnoses'. The preliminary search yielded 1407 abstracts.

We then combined each of the individual terms with mammography and restricted the search to clinical studies, reviews and meta-analyses published after 1996 and studies published in English. The search yielded 183 relevant abstracts.

The criteria for inclusion were all prospective or retrospective studies that evaluated: (1) the intervention as part of routine screening (with all cases included), and (2) studies that selected cancer cases and controls and then conducted evaluative study with number of experts. Studies in children were not considered. We included both studies that compared single reader + CADe with single user and studies that compared double readers with single reader + CADe.

These were independently reviewed by four authors for inclusion in the review. After review 52 abstracts were considered relevant and full reports were ordered. We also checked the reference list of all studies. Independent review identified 18 unique clinical evaluations, 12 reviews (including those with meta-analyses) and 7 editorials or responses to papers that were relevant.

We found one good quality systematic review (ref Taylor and Potts) addressing CADe mammography in screening and evaluated both included and excluded studies based on our search. Then, four authors independently abstracted the data on demographics, clinical characteristics and outcomes of the CADe mammography from all studies and reviews. Only explicit description of outcome events was tabulated. If the manuscript did not contain information then the endpoint was scored as missing.

III. Retrospective Reader Performance Studies

Alberdi et al ¹ conducted retrospective reader performance studies of the use of CADe mammography in the United Kingdom. The first study contained 30 normal cases, and 30 cancers, with 20 of these cancers missed by CADe in a clinical trial. These examinations were read by 20 readers from the clinical trial with the assistance of the R2 ImageChecker M1000 CADe device. The mean reader sensitivity in this study was only 52%. In the second study, 19 different readers read the examinations from the first study without CADe assistance. The investigators removed six cancers from the study set, suggesting that they were undetectable via mammography. The mean observed sensitivity for the remaining 24 cancers was 73% for readers who did not have CADe assistance and 61% for readers who were assisted by CADe.

Alberdi et al contended that, "...at least for some categories of cases, incorrect CADe output had a significant detrimental effect on human decisions in our studies." Their conclusion is greatly limited by their decision to employ different readers in the two studies. They assert, but do not demonstrate, that the readers were comparable in experience and professional background. Most importantly, it is well recognized that the sensitivity thresholds for different mammographic readers vary widely. Differences in sensitivity thresholds may have contributed to the results.

Taplin et al ² examined the effect of CADe on interpretive mammography performance. They began by identifying 527 cases of invasive carcinoma or ductal carcinoma in situ. stratified random sampling led to the inclusion of 114 case of cancer that occurred within 12 months of mammographic screening, 113 cases of cancer that occurred between 13 and 24 months of screening, and 114 screenings where cancer did not occur. The ImageChecker M2 1000 system, version 2.2 was used in the study. Each radiologist rated all of the mammographic examinations with and without CADe assistance in 2 independent sessions. Radiologists provided assessments that employed the BI-RADS coding system. For cancers that appeared within 1 year, the sensitivity with CADe without CADe was 63.2% and without CADe it was 62.0%. For cancers that occurred within 13-24 months of screening, the sensitivity without CADe was 33.5% and the corresponding sensitivity with CADe assistance was 32.3%. The observed differences were not statistically significant for either cancer group. Reported specificity with CADe assistance was 75% and 72% without CADe assistance. The effect of CADe assistance on specificity was not modified by breast density. In contrast to the majority of published CADe studies, the Taplin et al study found a decrease in sensitivity and an increase in specificity.

Brem et al ³ conducted a retrospective study of CADe performance in a data set of 177 missed breast cancers and 155 normal cases. The authors reported a calculated radiologist sensitivity of 75.4% without CADe and 91.4% with CADe. However, sequential reading with and without CADe did not occur. Brem et al noted that, "The benefit derived from the computer-aided detection system was calculated as proportional to the number of radiologists who correctly identified the lesions at blinded review." For the normal cases in the study, 1.0 mass and 0.25 microcalcification marks per mammogram were observed.

Butler et al ⁴ conducted a retrospective standalone study of CADe performance. The cases examined utilized were obtained from diagnostic mammograms. All 30 of the cases were unsuspected cancers that presented at a location away from the clinical finding that prompted the diagnostic mammogram. A total of 26 of the 30 cases were identified by the CADe. The standalone performance study is limited by the lack of assessment of the interaction between the radiologist and the CADe. Butler et al readily conceded, "In addition, we have no proof that a radiologist prospectively using the CADe system would take action on the basis of one of the CADe marks."

Warren et al ⁵ conducted a retrospective review of mammography records to determine the false-negative rate in screening mammography, the capability of computer-aided detection (CADe) to identify these missed lesions, and whether or not CADe increases the radiologists' recall rate. They reviewed all available mammograms that led to the detection of biopsy-proven cancer (n=1,083) and the most recent corresponding prior mammograms (n=427). Thirteen different facilities were included in this retrospective review. All the mammograms were originally

reviewed by a radiologist. For this study a panel of 5 radiologists evaluated the retrospectively visible prior mammograms by means of blinded review. Additionally, all mammograms were analyzed by a CADe system. The recall rates of 14 radiologists were prospectively measured before and after installation of the CADe system. The results show, the original radiologists' sensitivity was 79% (427 of [427 1 115]). At independent, blinded review by panels of radiologists, 27% (115 of 427) were interpreted as warranting recall on the basis of a statistical evaluation index; and the CADe system correctly marked 77% (89 of 115) of these cases. No statistically significant increase in the radiologists' recall rate was observed when comparing the values before (8.3%) and after (7.6%) installation of the CADe system. The authors conclude radiologists reading alone had a false-negative rate of 21% (115 of [427 1 115]). CADe prompting could have potentially helped reduce this false-negative rate by 77% (89 of 115) without an increase in the recall rate. Although the results of this study seem promising in favor of using CADe, the fact that the study was conducted in an enriched prevalence population (i.e. confirmed cancer cases) limits the generalizability of the results.

The study by Zheng et al ⁶ was conducted to assess the performance of radiologists in the detection of masses and microcalcification clusters on digitized mammograms, by using different computer-assisted detection (CADe) cuing environments. These researchers used 209 digitized mammograms depicting 57 verified masses and 38 microcalcification clusters in 85 positive and 35 negative cases. All records were interpreted independently by seven radiologists using five display modes. The first mode did not include CADe cuing. In all other modes suspicious regions identified with a CADe scheme were cued by using a combination of two cuing sensitivities (90% and 50%) and two false-positive rates (0.5 and 2.0 per image). Then receiver operating characteristic (ROC) study was performed by using soft-copy images. Results show that CADe cuing at 90% sensitivity and a rate of 0.5 false-positive region per image improved observer performance levels significantly ($P < 0.01$). As accuracy of CADe cuing decreased so did observer performances ($P < 0.01$). Cuing specificity affected mass detection more significantly, while cuing sensitivity affected detection of microcalcification clusters more significantly ($P < 0.01$). Reduction of cuing sensitivity and specificity significantly increased false-negative rates in non-cued areas ($P < 0.05$). Trends were consistent for all observers. These researchers conclude that CADe systems have the potential to significantly improve diagnostic performance in mammography. However, poorly performing schemes could adversely affect observer performance in both cued and non-cued areas.

Limitations of Retrospective Reader Performance Studies

There is an extensive array of limitations for retrospective reader performance studies of CADe mammography performance. First, in order to permit the time commitment of participating radiologists to be reasonable, the case mix is entirely unrealistic. In the actual practice of mammography screening only a tiny minority of examinations, less than one percent, will be true positives. The retrospective reader performance studies have true positive case mixes that vastly exceed this amount. The greater number of true positives can bias the results in a number of different ways. Radiologists will be aware of the far greater baseline probability of true positive examinations and may adjust their interpretations accordingly. In addition, there will be far fewer

false positive marks per true positive case for the radiologist to dismiss as the proportion of true negative cases will be far smaller than the proportion encountered in actual practice. Second, valuable patient data, such as a family history of breast cancer, were not provided in several retrospective reader performance studies. Third, access to and comparison with previous mammographic examinations, an integral component of the interpretative process in mammography, and indeed all radiographic examinations, is often not provided. Comparison with previous examinations can be instrumental in determining that a given abnormality is benign. Fourth, the retrospective reader performance setting does not fully reflect clinical reality. The profound human cost of failing to detect a cancer in practice is not found in this setting. Fifth, medicolegal considerations, that may heavily weigh on a radiologist in actual practice, are absent in this setting.

IV. Clinical Studies of CADe Mammography

Gilbert et al.⁷ determined whether the performance of a single reader using computer-aided detection (CADe) would match the performance achieved by two independent readers through a prospective equivalence trial with the matched comparison of single reader with CADe vs. double independent readers in UK. A total of 31,057 women seen during 2006-2007 at 3 mammography screening centers in UK were enrolled. Of them, 28,204 women received both single reading with CADe and double reading. A total of 227 cancers were detected, for an overall detection rate of 8.0/1000. The proportion of cancers detected was 199 of 227 (87.7%) for double reading and 198 of 227 (87.2%) for single reading with CADe ($P = 0.89$). The overall recall rates were 3.4% for double reading and 3.9% for single reading with CADe; the difference between the rates was small but significant ($P < 0.001$). The estimated sensitivity, specificity, and positive predictive value for single reading with CADe were 87.2%, 96.9%, 18.0% respectively; and were 87.7%, 97.4%, and 21.1% for double reading respectively. There were no significant differences between the pathological attributes of tumors detected by single reading with CADe and those of tumors detected by double reading alone. The authors concluded that single reading with CADe could be an alternative to double reading and could improve the rate of cancer detection from read by a single reader.

Fenton et al.⁸ determined the association between the use of CADe and the performance of screening mammography from 1998 through 2002 at 43 facilities of Mammography registry. They had data for 222,135 women (a total of 429,345 mammograms), including 2351 women who received a diagnosis of breast cancer within 1 year after screening. They calculated the specificity, sensitivity, and positive predictive value (PPV) of screening mammography with and without CADe, as well as the rates of biopsy and breast cancer detection and the overall accuracy, measured as the area under the ROC curve. Of note, only 7 facilities (16%) implemented CADe during the study period. For these 7 facilities, diagnostic specificity decreased from 90.2% before implementation to 87.2% after implementation ($p < 0.001$), the positive predictive value decreased from 4.1% to 3.2% ($p = 0.01$), and the rate of biopsy increased by 19.7% ($p < 0.001$). The increase in sensitivity from 80.4% before implementation of CADe to 84.0% after implementation was not significant ($p = 0.32$). The change in the cancer detection rate (including invasive breast cancers and ductal carcinomas in situ) was not significant (4.15 cases per 1000 screening mammograms before implementation and 4.20 cases after implementation, $p = 0.90$). Adjusted analyses of data from all 43 facilities showed that the

use of CADe was associated with significantly lower overall accuracy than was non-use (area under the ROC curve, 0.871 vs. 0.919; $p = 0.005$). The authors concluded that the use of CADe is associated with reduced accuracy of interpretation of screening mammograms. The increased rate of biopsy with the use of CADe is not clearly associated with improved detection of invasive breast cancer.

Gilbert et al.⁹ retrospectively determined if the use of a CADe can improve the performance of single reading of screening mammograms to match that of double reading in the UK. The study included a sample of 10,267 mammograms obtained in women aged 50 years or older who underwent routine screening at two breast screening centers in 1996. Mammograms that were double read in 1996 were randomly allocated to be re-read by 8 different radiologists using CADe. The cancer detection and recall rates from double reading and single reading with CADe were compared. The results showed that single reading with CADe led to a cancer detection rate that was significantly ($P=0.02$) higher than that achieved with double reading: 6.5% more cancers were detected by single reading with CADe than by double reading. However, the recall rate was higher for single reading with CADe than for double reading (8.6% vs. 6.5%, $P=0.001$). This was equivalent to relative increases of 15% and 32% in the cancer detection and recall rates, respectively. The authors concluded that single reading with CADe leads to an improved cancer detection rate and an increased recall.

Freer et al.¹⁰ prospectively assessed the effect of CADe on the interpretation of screening mammograms through a sequential clinical study in a community breast center. Over a 12-month period, 12,860 screening mammograms were interpreted with the assistance of a CADe system. Each mammogram was initially interpreted without the assistance of CADe, followed immediately by a reevaluation of areas marked by the CADe system. The results showed that, when comparing the radiologist's performance without CADe with that when CADe was used, there was (a) an increase in recall rate from 6.5% to 7.7%; (b) no change in the positive predictive value for biopsy at 38%; (c) a 19.5% increase in the number of cancers detected; and (d) an increase in the proportion of early-stage (0 and I) malignancies detected from 73% to 78%. The authors concluded that the use of CADe in the interpretation of screening mammograms can increase the detection of early-stage malignancies without undue effect on the recall rate or positive predictive value for biopsy.

Ko et al.¹¹ reported the results of CADe in a sequential study (without and subsequently with CADe) involving 5016 screening cases (45 cancers) read over a 26-month period. With the addition of CADe, the recall rate increased from 12% to 14%, while the cancer detection rate increased from 0.90% to 0.94%. Sensitivity of screening mammography with the use of CADe (94%) represented an absolute 4% increase over the sensitivity of the radiologist alone (90%). Specificity of screening mammography with and without the use of CADe was 99%. CADe detected two in situ cancers that were missed by the radiologist. The authors concluded that routine use of CADe significantly increases recall rates, while also increase the cancer detection rate and sensitivity by 4%-5%.

Gur et al.¹² reported the change in mammography recall and cancer detection rates after the introduction of a CADe in a clinical study involving 24 radiologists and 115,571 examinations. They found a 1.7% increase in cancer detection rate with CADe (3.55/1000 vs. 3.49/1000).

without CADe; $p = 0.68$). The recall rates were also similar for mammograms interpreted without and with CADe (11.39% vs. 11.40%, $P=0.96$). The subset analyses of 7 radiologists with high volume of mammograms reading demonstrated the similar results. The authors concluded that the introduction of CADe was not associated with statistically significant changes in recall and breast cancer detection rates.

Helvie, MA et al.¹³ evaluated a noncommercial CADe program for breast cancer detection with screening mammography through a sequential study on 2,389 patients' screening mammograms from 2 academic institutions in US. Thirteen radiologists who specialized in breast imaging participated in this study. For each case, the radiologist performed the first assessment without CADe. Then, the radiologist was shown CADe results and rendered a second assessment. Outcome included patients recall, biopsy, and 1-year follow-up examination. The results showed that 11 (0.46%) of 2,389 patients had mammographically detected nonpalpable breast cancers. Ten (91%) of 11 (95%CI 74%, 100%) cancers were correctly identified with CADe. Radiologist sensitivity without CADe was also 91% (10 of 11; 95%CI 74%, 100%). In 1,077 patients, 1-year follow-up results were available. Five (0.46%) patients developed cancers, of which the area where the cancers developed in 2 of these five patients was marked by the CADe in the preceding year. CADe also resulted in a 9.7% increase in recall rate from 14.4% to 15.8%. The authors concluded that the performance of the CADe program had a very high sensitivity of 91%.

V. Meta-Analyses and Narrative Reviews

Two meta-analyses^{14, 15} and one narrative review¹⁶ were found among the publications that resulted from our search criteria. Each meta-analysis included a different set of published studies because the questions that the authors aimed to answer were different for each of them. Each meta-analysis publication is first described followed by the narrative review and an assessment of the evidence is provided at the end of this section.

Meta-Analyses

The analysis performed by Taylor and Potts was conducted to assess how the use of CADe or double reading affects the cancer detection rate and the recall rate. Two sets of studies were reviewed: (1) studies comparing single reader versus single reader with CADe and (2) studies comparing double reading to single reading. They used the following inclusion criteria to select the studies:

1. *Types of studies*: Prospective and retrospective studies where the intervention was incorporated into routine screening work and all cases selected only on the basis of the usual screening criteria.
2. *Types of participants*: All studies of women in screening age (≥ 40)
3. *Types of interventions*: Studies using commercially available CADe systems and studies of double reading in which the second reading was performed by a trained reader but not a radiologist.
4. *Types of outcome measures*: Only studies reporting the impact on cancer detection rate and recall rate or studies from which these could be calculated.

If the comparisons were made based on the same mammogram the study was considered “matched”. When the performance of mammography in a facility was evaluated before and after introducing CADe, then the study was considered to be “unmatched”.

Single Reading versus CADe

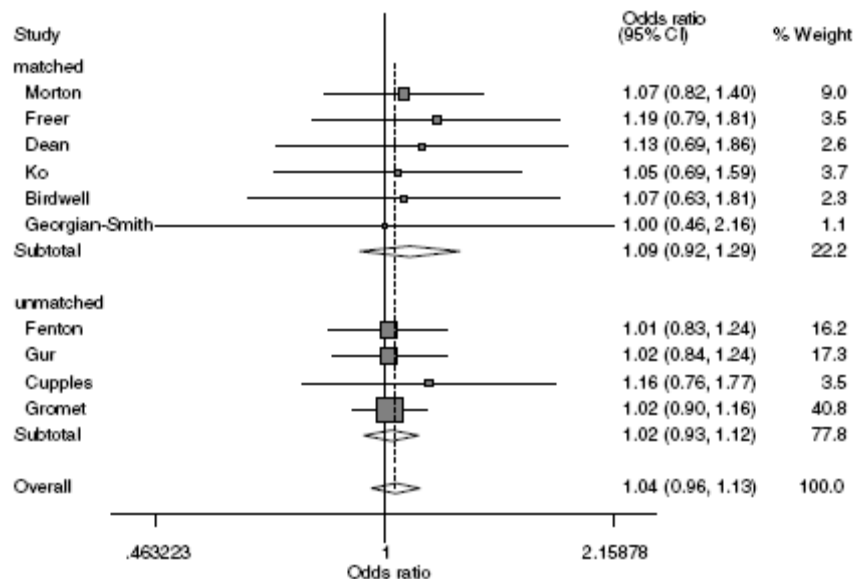
A total of 10 studies evaluating single reading versus single reader with CADe were included. There were six matched studies and 4 unmatched studies; all of them conducted in the USA. The proportional contribution to the cancer detection rate (CDR) and the recall rate were estimated as follows:

$$(\text{CDR}_{\text{CADe}} - \text{CDR}_{\text{Single reader/no CADe}}) / \text{CDR}_{\text{Single reader/no CADe}}$$

$$(\text{Recall Rate}_{\text{CADe}} - \text{Recall Rate}_{\text{Single reader/no CADe}}) / \text{Recall Rate}_{\text{Single reader/no CADe}}$$

The proportional contribution of CADe on CDR ranged from 0.00 to 0.20 and the proportional contribution of CADe on the recall rate ranged from 0.00 to 0.31.

The Odds Ratio for the effect of CADe on the CDR and the recall rate were estimated. Figure 1 shows the impact of CADe on the CDR. Each study is shown as a horizontal line. The length of the line indicates the width of the 95% confidence intervals. The position of the midpoint shows the measured effect. The size of the centre square reflects the contribution to the pooled estimates (largely determined by sample size). The summary results are shown as diamonds. The centre of the diamond shows the combined estimate of the effect, and the distance to the left and right extremities shows the 95% confidence interval.



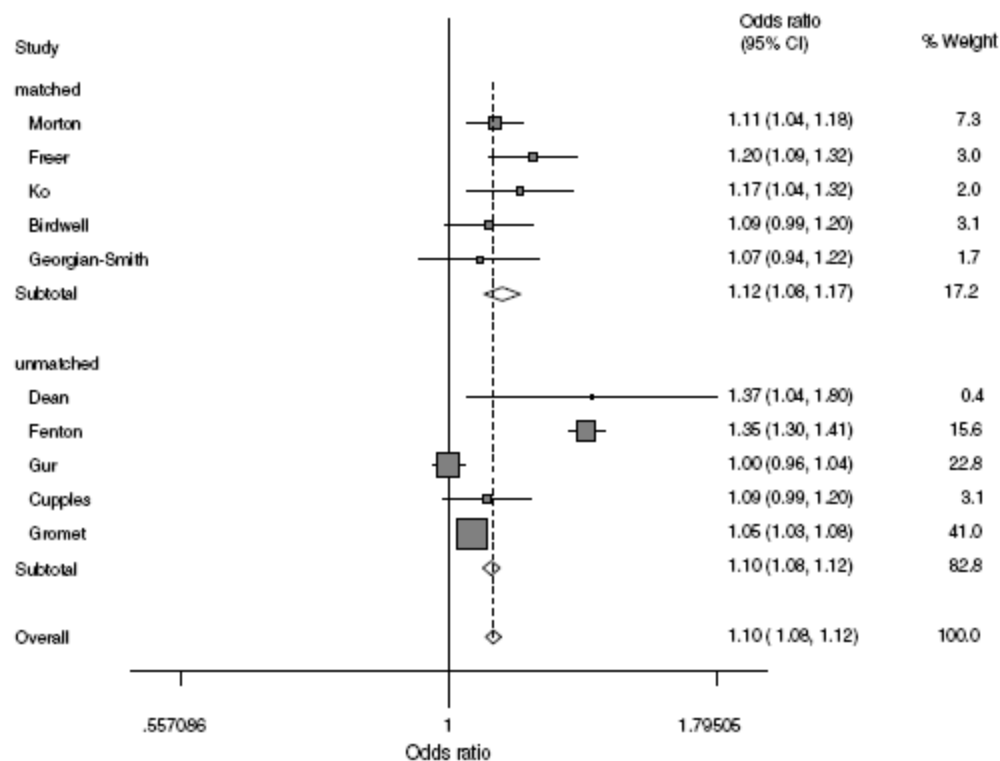
Source: Taylor and Potts, 2008; Page 802

Figure 1 Effect of CADe on Cancer Detection Rate

There is no evidence of heterogeneity between or within matched and unmatched studies.

Estimates for individual studies as well as the overall estimate were not statistically significant, meaning there is no significant increase in the CDR with the use of CADe. The results were similar when the study by Fenton⁸ was excluded.

The results on the effect of CADe on the recall rate are presented in Figure 2 (below). All the studies show increased recall rates, but there is evidence of heterogeneity (overall test, $p < 0.001$). The matched studies do not show heterogeneity, but the unmatched studies do ($p < 0.001$). A significant heterogeneity test result is still observed if either the study by Fenton⁸ or Gur¹² are included; however the other papers are consistent. The marked difference between these two studies is not explained by this analysis. The overall estimates are statistically significant; but the best estimate for the effect on the recall rate is from the matched studies, which shows a significant 12% increase in the recall rate.



Source: Taylor and Potts, 2008; page 803

Figure 2 Effect of CADe on Recall Rate

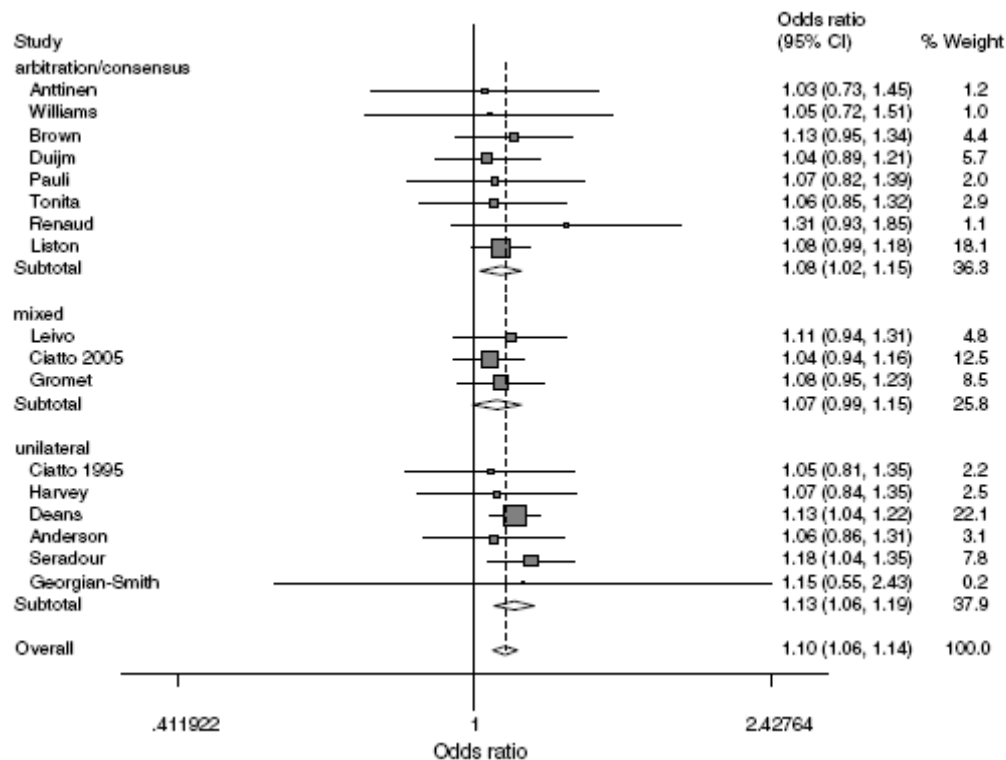
Single Reading versus Double Reading

There were 17 studies that compared single reading with double reading, all used a matched design. In many countries it is standard practice for each screening mammogram to be viewed by two readers who either confer on discordant cases or refer them to arbitration. Five of the 17 studies included in the meta-analysis used arbitration, 3 used consensus, 3 were mixed and 6 were unilateral. The studies were conducted in France (2), UK (5), Canada (1), USA (3), Netherlands (1), Finland (2), New Zealand (1) and Italy (2).

The proportional impact of double reading on the CDR ranged from 0.03 to 0.31; and the

proportional impact of double reading on the recall rate ranged from -0.39 to 0.38.

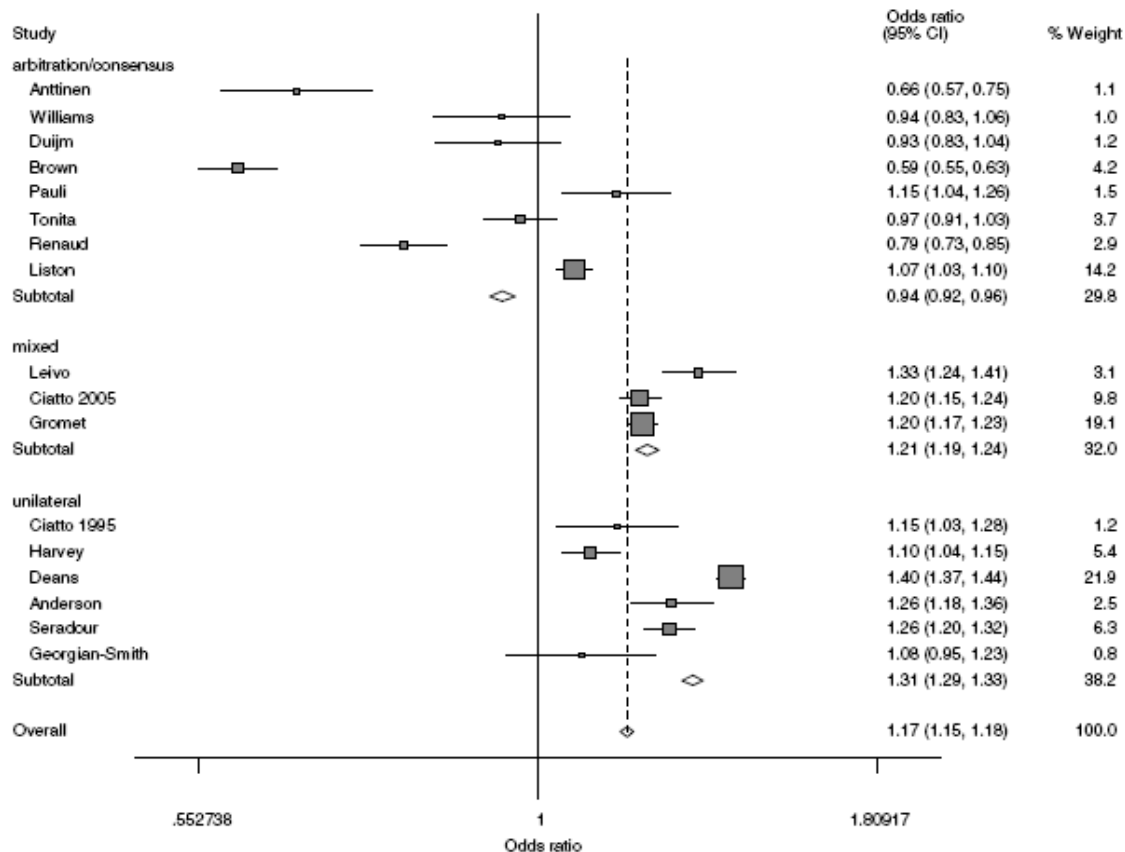
Figure 3 (below) presents the impact of double reading on the CDR. There is no evidence of heterogeneity. Most of the individual estimates show increase cancer detection rates; but, not statistically significant. However, the overall estimate shows a significant 10% increase in the cancer detection rate with the use of double reading.



Source: Taylor and Potts, 2008; page 803

Figure 3 Effect of Double Reading on Cancer detection Rate

Figure 4 (below) presents the effect of double reading on the recall rate. The evidence for the effect of double reading on recall rate is not as clear as there is heterogeneity between and within the groups (arbitration/consensus, mixed studies and unilateral). All the mixed and unilateral studies show increases in recall rate. However, arbitration studies show a decrease (overall estimate); but two of them, including one of the largest studies,¹⁷ show a significant increase.



Source: Taylor and Potts, 2008; page 804

Figure 4 Effect of Double Reading on Recall Rate

The analysis by Noble and colleagues was performed to assess the diagnostic performance of CADe for screening mammography in terms of sensitivity, specificity, recall rate, biopsy rate and cancer diagnosis rate¹⁵. The inclusion criteria used consisted of:

1. Only full-length (no poster presentations or abstracts) English-language publications that enrolled an asymptomatic population of at least 10 women undergoing plain-film mammography for routine breast cancer screening, that reported original data.
2. Studies with mixed screening and diagnostics were only be included if at least 85% of the population was asymptomatic or if findings for asymptomatic women were reported separately.
3. Studies could be prospective or retrospective, but must have enrolled patients randomly or consecutively.
4. Studies must have reported data to enable calculation of sensitivity and specificity, cancer diagnosis rate, recall rate and biopsy rate.

Their first search resulted in 71 potential relevant studies. Forty four publications were excluded because did not enroll patients randomly or consecutively. The other 24 publications were excluded mostly due to not enrolling screening population or enrolling fewer than 85% of

patients for screening and not reporting screening data separately from diagnostic data. Some were excluded due to lack of data necessary to compute the endpoints of interest for the meta-analysis.

Three of the excluded studies were analyzed in the meta-analysis by Taylor and Potts^{12, 14, 18, 19}. The studies by Gur¹² and Cupples¹⁸ were excluded from this analysis because they did not provide one year follow-up data to confirm negative findings and did not provide the data needed for the endpoints of interest; and the study by Dean¹⁹ was excluded because of the high prevalence of women seeking mammography for reasons other than screening (about 40%).

Seven publications were included in this analysis; three were retrospective studies, four were prospective studies.

Sensitivity and Specificity

Three studies reported sensitivity and specificity for single reader with CAdE, with reference to cancer diagnosis one year later. Sensitivities of each study were: 72.2%²⁰, 84.0%⁸ and 90.4%²¹. The overall sensitivity is estimated as 86% (95%CI 84.2, 87.6).

Specificities were: 92.3%²⁰, 87.2%⁸ and 89.7%²¹. The overall specificity is estimated as 88.2% (95%CI 88.1, 88.3).

Incremental Cancer Detection

Four studies provided data for this outcome^{10, 11, 22, 23}. In these studies a single radiologist assessed the mammogram; then the radiologist re-evaluated the mammogram with the help of CAdE. The incremental cancer detection was estimated as 50 additional women per 100,000 screened (95%CI 30, 80). This estimate was robust to sensitivity analysis, but is not precise (wide CI). Of all the women who were recalled based on CAdE findings, 4.1% (95%CI 2.7, 6.3) were diagnosed with cancer. The data from these studies were not heterogeneous ($p < 0.001$); and the meta-analysis was robust to sensitivity analysis.

False Positives

The same four studies provided data to assess the false positive outcome. There was evidence of heterogeneity (p value not provided). The analysis for incremental recall of healthy women show 1,190 (95%CI 1,090, 1,290) additional healthy women per 100,000 women screened were recalled on the basis of CAdE results. Ninety six percent of women recalled by CAdE were not diagnosed with cancer (95%CI: 93.0, 97.3%). These findings were not substantially heterogeneous, and the meta-analysis was robust to sensitivity analysis.

The incremental biopsy rate of healthy women is 80 (95%CI 60, 110) additional women per 100,000 screened. Substantial heterogeneity was not detected and the analysis was robust.

Sixty five percent (95%CI 52.3, 76.0%) of the women who underwent biopsy based on CAdE results were healthy. The remaining 35.9% (95%CI 24.7, 48.9%) of women were diagnosed with cancer. There was no evidence of heterogeneity and the estimate was robust to sensitivity

analysis.

Narrative Review

Helvie's paper¹⁶ lacks details about how the studies that were evaluated were selected for inclusion in the analysis. The author first discusses several factors that are known to influence the radiologist performance, like: radiologist expertise, association between sensitivity and recall rate, and observation time and sensitivity. He argues these factors affect reader's performance independent from double reading or CADe use.

Helvie states reader variability is the weakest link in the imaging change, and then proceeds to discuss several papers that deal with readers' expertise and performance. Parameters that have been associated with reader's performance are: higher volume, fellowship training, continuing medical education and regular participation in radiologic-pathologic correlation conference with case-specific outcomes feedback¹⁶.

He also discusses how increasing the recall rate and/or the false positives can improve sensitivity, independent of double reading or CADe. To support this statement he discusses studies by Yankaskas and colleagues²⁴ and Otten and colleagues²⁵. In the study by Yankaskas in North Carolina, sensitivity improved as the recall rate increased; and in the study by Otten increases in the false positive rate increased the sensitivity of readers' performance. Also discussed the publication by Gur²⁶, which found on average a 0.22/1000 cancer detection rate improvement occurred for every 1% absolute increase in the recall rate.

Observation time can also improve sensitivity. Helvie discusses 4 studies to support this statement²⁷⁻³⁰. He concludes that although a quick reading of a mammographic image by an experienced radiologist detects most malignancies, a minority of cases may be overlooked.

CADe Review

CADe systems are programmed at certain level of sensitivity and specificity (with strong emphasis on sensitivity). The settings can be adjusted depending on the desire of the customer. Helvie discusses several retrospective studies that show using CADe improved detection; improvement ranged from 21.2% to 77%^{3, 5, 31}. However, he also discusses some of the set backs of CADe use. In one study the sensitivity for CADe was 50% compared to 59% of the expert³². Another study show a non-significant 4.2% increase in sensitivity with a significant increased in the recall rate of 44%³³.

Helvie also included eight clinical trials of screening mammography with CADe in the USA^{10-13, 18, 19, 22, 23}. All show increased detection rates as well as increased recall rates. The change in detection rate ranged from 1.7% to 19.5%; and the change in recall rate ranged from 0.1% to 26%.

Double Reading Review

Nine studies were included on double reading^{2, 34-40}. We were not able to find the reference for

the study by Seradour, B and colleagues (1996). The studies were conducted in Finland, Scotland, Sweden, Italy, UK, France and USA (3). Two of the studies used consensus, three were independent reviewers, two independent reviewer/discussion/expert, and two were experimental. The final recall rate ranged from 2.5% to 14%, the biopsy PPV ranged from 21% to 80%, the detection rate ranged from 4.3/1000 to 6.7/1000, the detection rate change ranged from 4.6% to 15% and the recall rate change ranged from -45% to +45%.

Assessment of Evidence Provided by Meta-Analyses and Narrative Review

Although the two meta-analyses were designed to address different questions, both come-up with fundamentally similar conclusions.

Taylor and Potts conclude that: (1) their pooled estimates suggest that CADe may change the threshold for recalls rather than improve the accuracy of screening; (2) CADe impact is diminished by the high number of false positive prompts; (3) CADe increases recall rate; (4) Not enough evidence to conclude CADe improves cancer detection rate; (5) there is evidence that double reading increases cancer detection rates; and (5) double reading with arbitration can lower the recall rates. They recommend: (1) CADe developers should improve specificity; and (2) further research to evaluate heterogeneity in reported recall rates.

This meta-analysis was well designed with clearly stated inclusion criteria for the studies that were selected and the endpoints of interest (cancer detection rates and recall rates). The statistical analyses performed are valid and commonly used in meta-analysis. They considered the studies that have previously created some controversy by Fenton⁸ and Gur¹², which were driving heterogeneity in recall rates. Sensitivity analyses show their overall estimates remain the same including or excluding those studies.

Noble, et. al. used different inclusion criteria and different study endpoints (sensitivity, specificity, recall rate, biopsy rate and cancer diagnosis rate). These determined the inclusion or exclusion of certain studies. However, the authors did a good job explaining the reasons for excluding studies from the analysis. Despite the differences in methodology (compared to Taylor and Potts), similar findings were reported. These authors conclude: (1) CADe increases recall rate of healthy women; (2) increases biopsy rate of healthy women; (3) limited impact: use of CADe will identify 50 additional cases in 100, 000 screened women. Although they believe their results are robust, they recommend frequent monitoring of the literature for this type of technology.

This meta-analysis was also well designed with clearly stated objectives, inclusion/exclusion criteria and the endpoints of interest. Although, charts common to meta-analysis are not provided, the statistical analysis was valid and commonly used for this type of analysis.

The narrative review by Helvie was well written and seems to cover a wide range of publications. However, the review does not include the methodology used for the literature search, the inclusion and exclusion criteria used to select the studies and no formal meta-analysis was performed. Due to differences in the methodologies of the studies that were included, the

estimates are not comparable. The review does, however, provide valuable input on the factors related with performance of radiologist (independent from CADe and double reading) and provides a narrative of the results of studies that evaluated performance of CADe and double reading.

VI. CADe Mammography: Research Recommendations for the Future

This memo has reviewed a number of research studies, meta-analyses and review papers conducted to evaluate the safety and effectiveness of incorporating CADe readers in the interpretation of mammography exams. All currently approved CADe devices are labeled for use as “second readers” in both routine screening and diagnostic mammograms. For use in diagnostic mammography, CADe devices are only approved for symptomatic patients with standard mammographic views. We are not aware of other situations where mammography CADe is used. These studies do not present a clear and reproducible set of results that clearly support the safety and effectiveness in real-world situations, leading to conflicting recommendations regarding the incorporation of this technology into clinical practice. Although there is evidence that double-reading increases cancer detection rates and that double-reading with arbitration lowers the recall rate¹⁴, the findings using CADe as the second reader are mixed. Taylor and Potts stress that changes noted in CADe may be explained by change in the threshold for the threshold for the recall rate (since the recall rate is increased) rather than improvement in the accuracy of screening. This concern raises an area of question that must be evaluated in future studies.

To date, there are no randomized clinical trials that compare single-reader or double-reader to single reader plus CADe for the evaluation of mammography scans. A future three-arm randomized controlled trial that compares the effectiveness of single reader, double-reader and single-reader plus CADe for screening mammography may resolve the problems in a consistent set of findings in clinical studies. The inclusion of double-reader is important since studies have found improvements in sensitivity with a decrease in recall rates; however specificity also decreased¹⁶. Gilbert et al⁷ compared double-reader to single-reader plus CADe and found that there was no difference in sensitivity, however there was a small but significant difference in recall rates.

Any study must be designed to collect information on the associated cofactors that may affect findings. These factors are listed in Table 1.

Table 1: Study Variables Necessary for Full Evaluation

Factor	Reasons for Inclusion
Race	Breast density and cancer types and rates vary across different race and ethnic groups
Breast Density	Breast density affects sensitivity. Ho et al (2003) grouped women based on ⁴¹ breast density scores and found a marked decrease in sensitivity as breast density increased
Radiology Training and Experience with mammogram	Helvie ¹⁶ (see also Elmore et al ⁴² and Robinson ⁴³ notes that the weakest link in the imaging chain is reader interpretive variability. Factors that affect interpretation include experience,

Factor	Reasons for Inclusion
reading	Board Certification/fellowship training, volume and observation time
Radiologist Decision-Making	The impact of the medicolegal environment on decision-making should be evaluated including the decision-making to discount a mark made by CADe systems.
CADe Training	Due to the differences and complexities of decision-making using CADe systems, a detailed training program is necessary to improve results. Any comparative study should outline specifically the type and duration of training.
Types of cancer detected	CADe systems are more effective in identifying microcalcification compared to masses, particularly in dense breasts.
Baseline breast cancer prevalence in the population	When evaluating sample size, baseline prevalence should be used in the power analysis.
Recall rate	There is a positive relationship between recall rate and sensitivity
Observation time	Helvie ¹⁶ notes that some cancer types, or the position of the cancer are only detected if the observation time is sufficient to study the film.
Adjustments to the CADe programs	CADe systems can be adjusted to balance sensitivity and specificity by adjusting the number of marks per film. This adjustment will affect any study and should be included in the evaluation of the system.
Single-Reader + CADe versus two independent readers one with CADe and one without CADe	When a study is designed using a single reader who will look and evaluate the film, make a determination and then look again with CADe and make a determination, there is the potential that both readings are biased by the knowledge that there will be a CADe available. This may cause the reader to become more conservative in the non-CADe assessment, affecting comparisons of sensitivity and specificity.
Biopsy Rate	The goal of an efficient screening tool is to minimize unnecessary exposure to additional intervention. False-negative readings often lead to biopsy. This procedure poses an additional set of both physical and psychological risks as well as additional health care costs.
Baseline sensitivity/PPV, specificity/NPV and recall rate	In order to fully evaluate the impact of double-reader or reader + CADe, it is critical to have baseline information on performance in the environment where the study is conducted
Study PPV and NPV	Although sensitivity analysis will predict the ability of the reading method to identify cancers, the true ability of the test is measured by both positive predictive value and negative predictive values since they include population parameters. Additionally, these values should be calculated on all screening cases, often they are only calculated on identified cancer cases, this may provide

Factor	Reasons for Inclusion
	misleading information for interpretation.
Area under the curve or ROC analysis	Since there is a positive association between recall rate and sensitivity, a more accurate measure of improvement through CADe would be ROC analysis that would give the absolute improvement by controlling for other factors such as a change in threshold
Population characteristics <ul style="list-style-type: none"> • Previous Screens • Estrogen Use 	Patient history can affect cancer rates, the decision-making of the clinician. Estrogen can affect breast density.
Patient Safety <ul style="list-style-type: none"> • Pain due to compression • Anxiety • Impact of increased recall rate • False-negative rate • Number of missed cancer cases 	Potential unintended adverse consequences from the use of mammography CADes in actual practice is a critical consideration and should be explored as secondary endpoints through administration of tools for pain and anxiety and calculation of increased recall rates, additional biopsies, the impact of the false-negative rate on both patients and cost and the impact of the number of missed cancer cases.
Patient Survival	Available studies are limited by little information on long-term survival following screening. Since the goal of mammography is early breast cancer detection to prevent breast cancer death, analyses should include long-term follow-up to determine if women develop breast cancer (i.e. cancer was missed at screening or detection was not early) and subsequently die due to the disease.

In summary, the benefits of mammography CADe must be weighed against the potential harms. The assessment of mammography CADe complicated by the absence of a clear and acceptable trade-off between increased recall and negative biopsy rates and increased sensitivity in the detection of breast cancer. Increased recall and negative biopsy rates are intrinsically undesirable on the grounds of cost, morbidity, and psychological stress. This is a value judgment and a largely philosophical and not quantitative question. The potential negative impact of medicolegal considerations on mammography CADe performance during actual conditions of use must not be underestimated.

Research indicates that addition of a second reader improves sensitivity and decreases the recall rate; however, when CADe is used in place of the second reader, findings on its safety and effectiveness are mixed. There is no objective number of recalls and negative biopsies that offset the detection of one breast cancer. Currently there are no randomized clinical trials that have

compared the different reader approaches, nor is there a consistency in data collected in observational studies to allow extensive comparison. Additionally, studies using test sets often lack sufficient conditions to simulate real-life settings which limits the ability to generalize findings to collect the information in Table 1, including long-term follow-up may be a preferred option due to cost considerations.

Awareness of the potential unintended adverse consequences from the use of mammography CADes in actual practice is a critical consideration. Helvie advances a cogent argument that the use of the CADe may create the potential for a decrease in the effort of the radiologist in examining the mammogram, diminishing the benefits of CADe utilization.

Newer modalities such as MRI and ultrasound also show potential; however these modalities have been used as diagnostic, not screening procedures and have not been sufficiently tested for screening purposes. It is essential that further evaluation which includes all elements listed in Table 1 be conducted on large, diverse samples of women in clinical settings. Although the randomized controlled trial is preferred, large observational studies have merit as well.

VII. Conclusions

Retrospective reader performance studies support the contention that CADe mammography devices can detect an appreciable proportion of breast cancers, including those cancers missed by radiologists. However, clinical studies provide a more nuanced picture of CADe performance. The data in the aggregate from these studies and published meta-analyses point to an increase in the recall rate and biopsy rate emanating from CADe usage. This is not surprising, given that in the actual mammographic screening environment the number of false-positive marks provided by the CADe will greatly dwarf the number of true-positive marks. Individual clinical studies of CADe use in actual practice in general suggest increased sensitivity with CADe use. The meta-analysis published by Noble et al is largely supportive of CADe usage as it concludes that use of CADe will identify 50 additional breast cancer cases in 100,000 screened women. The meta-analysis conducted by Potts is less sanguine, as it argues that there is not yet enough evidence to conclude that CADe improves the breast cancer detection rate.

The continued refinement of CADe algorithms is warranted with the goal of improving specificity and decreasing the number of false-positive marks radiologists must contend with. The most striking limitation of the available data is the absence of any data on the effect on patient survival from CADe usage. Large, prospective trials of CADe use that examine its effect on patient survival would be highly beneficial.

CADe Mammography Review Team
CDRH/OSB/DEPI

References

1. Alberdi E, Povyakalo AA, Strigini L, et al. Use of computer-aided detection (CADe) tools in screening mammography: a multidisciplinary investigation. *Br J Radiol.* 2005;78 Spec No 1:S31-40.
2. Taplin SH, Rutter CM, Elmore JG, Seger D, White D, Brenner RJ. Accuracy of screening mammography using single versus independent double interpretation. *AJR Am J Roentgenol.* May 2000;174(5):1257-1262.
3. Brem RF, Baum J, Lechner M, et al. Improvement in sensitivity of screening mammography with computer-aided detection: a multiinstitutional trial. *AJR Am J Roentgenol.* Sep 2003;181(3):687-693.
4. Butler SA, Gabbay RJ, Kass DA, Siedler DE, O'Shaughnessy K F, Castellino RA. Computer-aided detection in diagnostic mammography: detection of clinically unsuspected cancers. *AJR Am J Roentgenol.* Nov 2004;183(5):1511-1515.
5. Warren Burhenne LJ, Wood SA, D'Orsi CJ, et al. Potential contribution of computer-aided detection to the sensitivity of screening mammography. *Radiology.* May 2000;215(2):554-562.
6. Zheng B, Ganott MA, Britton CA, et al. Soft-copy mammographic readings with different computer-assisted detection cuing environments: preliminary findings. *Radiology.* Dec 2001;221(3):633-640.
7. Gilbert FJ, Astley SM, Gillan MG, et al. Single reading with computer-aided detection for screening mammography. *N Engl J Med.* Oct 16 2008;359(16):1675-1684.
8. Fenton JJ, Taplin SH, Carney PA, et al. Influence of computer-aided detection on performance of screening mammography. *N Engl J Med.* Apr 5 2007;356(14):1399-1409.
9. Gilbert FJ, Astley SM, McGee MA, et al. Single reading with computer-aided detection and double reading of screening mammograms in the United Kingdom National Breast Screening Program. *Radiology.* Oct 2006;241(1):47-53.
10. Freer TW, Ullissey MJ. Screening mammography with computer-aided detection: prospective study of 12,860 patients in a community breast center. *Radiology.* Sep 2001;220(3):781-786.
11. Ko JM, Nicholas MJ, Mendel JB, Slanetz PJ. Prospective assessment of computer-aided detection in interpretation of screening mammography. *AJR Am J Roentgenol.* Dec 2006;187(6):1483-1491.
12. Gur D, Sumkin JH, Rockette HE, et al. Changes in breast cancer detection and mammography recall rates after the introduction of a computer-aided detection system. *J Natl Cancer Inst.* Feb 4 2004;96(3):185-190.
13. Helvie MA, Hadjiiski L, Makariou E, et al. Sensitivity of noncommercial computer-aided detection system for mammographic breast cancer detection: pilot clinical trial. *Radiology.* Apr 2004;231(1):208-214.
14. Taylor P, Potts HW. Computer aids and human second reading as interventions in screening mammography: two systematic reviews to compare effects on cancer detection and recall rate. *Eur J Cancer.* Apr 2008;44(6):798-807.
15. Noble M, Bruening W, Uhl S, Schoelles K. Computer-aided detection mammography for breast cancer screening: systematic review and meta-analysis. *Arch Gynecol Obstet.* Nov 21 2008.

16. Helvie M. Improving mammographic interpretation: double reading and computer-aided diagnosis. *Radiol Clin North Am.* Sep 2007;45(5):801-811, vi.
17. Liston JC, Dall BJ. Can the NHS Breast Screening Programme afford not to double read screening mammograms? *Clin Radiol.* Jun 2003;58(6):474-477.
18. Cupples TE, Cunningham JE, Reynolds JC. Impact of computer-aided detection in a regional screening mammography program. *AJR Am J Roentgenol.* Oct 2005;185(4):944-950.
19. Dean JC, Ilvento CC. Improved cancer detection using computer-aided detection with diagnostic and screening mammography: prospective study of 104 cancers. *AJR Am J Roentgenol.* Jul 2006;187(1):20-28.
20. Georgian-Smith D, Moore RH, Halpern E, et al. Blinded comparison of computer-aided detection with human second reading in screening mammography. *AJR Am J Roentgenol.* Nov 2007;189(5):1135-1141.
21. Gromet M. Comparison of computer-aided detection to double reading of screening mammograms: review of 231,221 mammograms. *AJR Am J Roentgenol.* Apr 2008;190(4):854-859.
22. Birdwell RL, Bandodkar P, Ikeda DM. Computer-aided detection with screening mammography in a university hospital setting. *Radiology.* Aug 2005;236(2):451-457.
23. Morton MJ, Whaley DH, Brandt KR, Amrami KK. Screening mammograms: interpretation with computer-aided detection--prospective evaluation. *Radiology.* May 2006;239(2):375-383.
24. Yankaskas BC, Cleveland RJ, Schell MJ, Kozar R. Association of recall rates with sensitivity and positive predictive values of screening mammography. *AJR Am J Roentgenol.* Sep 2001;177(3):543-549.
25. Otten JD, Karssemeijer N, Hendriks JH, et al. Effect of recall rate on earlier screen detection of breast cancers based on the Dutch performance indicators. *J Natl Cancer Inst.* May 18 2005;97(10):748-754.
26. Gur D, Sumkin JH, Hardesty LA, et al. Recall and detection rates in screening mammography. *Cancer.* Apr 15 2004;100(8):1590-1594.
27. Christensen EE, Murry RC, Holland K, Reynolds J, Landay MJ, Moore JG. The effect of search time on perception. *Radiology.* Feb 1981;138(2):361-365.
28. Krupinski EA. Visual search of mammographic images: influence of lesion subtlety. *Acad Radiol.* Aug 2005;12(8):965-969.
29. Mello-Thoms C, Hardesty L, Sumkin J, et al. Effects of lesion conspicuity on visual search in mammogram reading. *Acad Radiol.* Jul 2005;12(7):830-840.
30. Nodine CF, Mello-Thoms C, Kundel HL, Weinstein SP. Time course of perception and decision making during mammographic interpretation. *AJR Am J Roentgenol.* Oct 2002;179(4):917-923.
31. Ikeda DM, Birdwell RL, O'Shaughnessy KF, Sickles EA, Brenner RJ. Computer-aided detection output on 172 subtle findings on normal mammograms previously obtained in women with breast cancer detected at follow-up screening mammography. *Radiology.* Mar 2004;230(3):811-819.
32. Thurfjell E, Thurfjell MG, Egge E, Bjurstam N. Sensitivity and specificity of computer-assisted breast cancer detection in mammography screening. *Acta Radiol.* Jul 1998;39(4):384-388.
33. Ciatto S, Del Turco MR, Risso G, et al. Comparison of standard reading and computer

- aided detection (CADe) on a national proficiency test of screening mammography. *Eur J Radiol.* Feb 2003;45(2):135-138.
34. Anderson ED, Muir BB, Walsh JS, Kirkpatrick AE. The efficacy of double reading mammograms in breast screening. *Clin Radiol.* Apr 1994;49(4):248-251.
 35. Anttinen I, Pamilo M, Soiva M, Roiha M. Double reading of mammography screening films--one radiologist or two? *Clin Radiol.* Dec 1993;48(6):414-421.
 36. Beam CA, Sullivan DC, Layde PM. Effect of human variability on independent double reading in screening mammography. *Acad Radiol.* Nov 1996;3(11):891-897.
 37. Ciatto S, Del Turco MR, Morrone D, et al. Independent double reading of screening mammograms. *J Med Screen.* 1995;2(2):99-101.
 38. Harvey SC, Geller B, Oppenheimer RG, Pinet M, Riddell L, Garra B. Increase in cancer detection and recall rates with independent double interpretation of screening mammography. *AJR Am J Roentgenol.* May 2003;180(5):1461-1467.
 39. Thurfjell EL, Lernevall KA, Taube AA. Benefit of independent double reading in a population-based mammography screening program. *Radiology.* Apr 1994;191(1):241-244.
 40. Warren RM, Duffy SW. Comparison of single reading with double reading of mammograms, and change in effectiveness with experience. *Br J Radiol.* Sep 1995;68(813):958-962.
 41. Ho WT, Lam PW. Clinical performance of computer-assisted detection (CADe system in detecting carcinoma in breasts of different densities. *Clin Radiol.* Feb 2003;58(2):133-136.
 42. Elmore JG, Wells CK, Lee CH, Howard DH, Feinstein AR. Variability in radiologists' interpretations of mammograms. *N Engl J Med.* Dec 1 1994;331(22):1493-1499.
 43. Robinson PJ. Radiology's Achilles' heel: error and variation in the interpretation of the Rontgen image. *Br J Radiol.* Nov 1997;70(839):1085-1098.